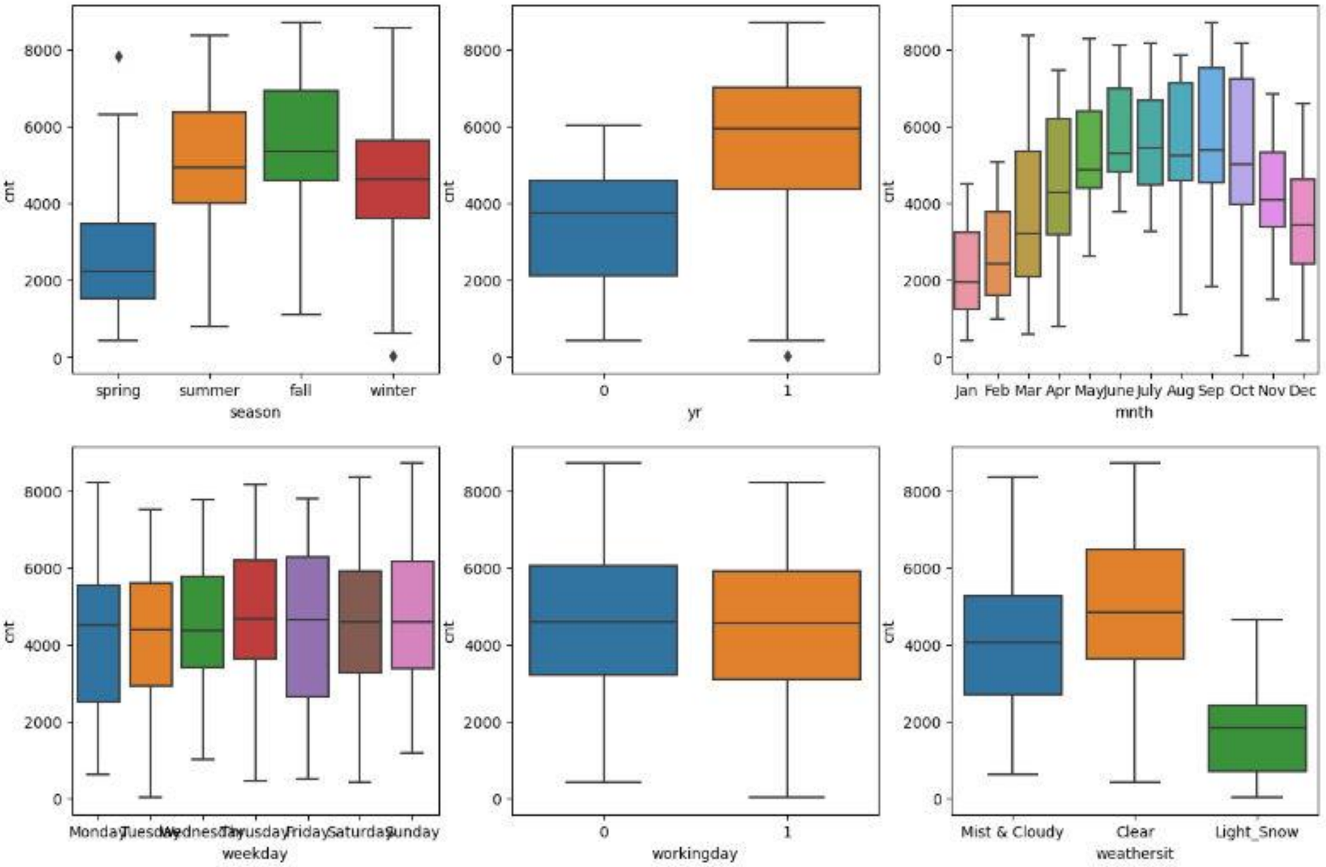
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
ANSWER:
Season, Yr, Mnth, Holiday, Weekday, Weathersit are categorical variables in the dataset.
From the analysis, it can be inferred that:
-Season :- The FALL season has the highest demand of rental bikes
-Year:- Demand has grown for the next year i.e 2019
- mnth:- demand has continuously grown till june and there is fall from sept till dec
- weekday friday has most demand however no proper inference can be taken at this point
- The clear weathersit has highest demand.



2. Why is it important to use drop_first=True during dummy variable creation?
ANSWER:
By using drop_first=True we intend to drop the first level of dummy variable created. By using drop_first=True we avoid the most vital part of dummy creation that is multicollinearity.
We use Syntax of dummy=pd.get_dummies
Lets say we have 3 types of values in categorical column example as shown below

Travel Type	Train	Bike	Plane
Plane	0	0	1
Bike	0	1	0

Train	1	0	0

3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

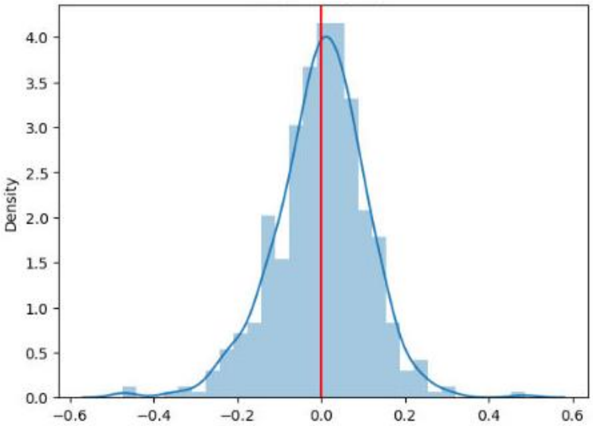
ANSWER:

As per the pair plot –the highest correlation with target variable is clearly “temp” and “atemp”.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

ANSWER:

The following are the ways to validate the assumptions of linear Regressions
1.The independent variables should not be highly correlated with each other
2. The error or residual analysis should follow normal distribution . This is done by ploting x-y plot of residual and further visualizing them
3. The relationship between the independent variables and the target variable should be linear
4. The residuals (the differences between the predicted and actual values) have constant variance (homoscedasticity)



5.Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

ANSWER:

Based on the final model the top three features based on coefficient are the variables are **“yr “** : Positive Coef **, “temp”** : Positive Coef and “**weathersit_Light_Snow”**: Negative Coef are the three features contributing significantly towards explaining the demand of the shared bike.

Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares.

The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

In the example above, y is the dependent variable, and x_1, x_2, and so on, are the explanatory variables. The coefficients (b_1, b_2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b_0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. What is Pearson’s R?

Ans: In statistics, the **Pearson correlation coefficient** (**PCC**, pronounced /ˈpɪərsən/) **—** also known as **Pearson's *r***, the **Pearson product-moment correlation coefficient** (**PPMCC**), the **bivariate correlation**,[1] or colloquially simply as **the correlation coefficient**[2] **—** is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is the process of normalizing the range of features in a dataset. Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

In both cases, you're transforming the values of numeric variables so that the transformed data points have specific helpful properties. The difference is that: in scaling, you're changing the range of your data, while. in normalization, you're changing the shape of the distribution of your data.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The VIF indicates how much collinearity has increased the variance of the coefficient estimate.(VIF) is equal to 1/(1-Ri^2).VIF=infinity if there is perfect correlation.Where R-1 denotes the R-square value of the independent variable for which we want to see how well it is explained by other independent variables.If an independent variable can be completely described by other independent variables. It has perfect correlation and has an R-squared value of 1.As a result , VIF = 1/(1-1) provides VIF = 1/0,which is infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.